Lecture 6
Linkage analysis

Claus Ekstrøm. Aug 2004. Royal Veterinary and Agricultural University, Denmark.

At the time of meiosis (i.e., when a reproductive cell divides) each chromosome duplicates to form two sister strands, or chromatids, joined together at the centromere region (see Figure 1). The homologous chromosomes form pairs, creating a tetrad consisting of the four chromatids. Non-sister chromatids then adhere to each other at various regions called chiasmata, where portions of the chromatids of one homologue are broken and exchanged with corresponding portions of the chromatids of the other homologue. The breaks of these crossovers are resealed, and the result is that the sister chromatids of a single homologue no longer contain identical genetic material. The maternal homologue now contains portions of the paternal homologue and vice versa. Thus, crossing over is an important mechanism for recombining the genetic material from the two parents — so important that observations of meioses suggest that almost always at least one chiasma occurs on each chromosome resulting in one obligatory crossover.

Figure 1: Crossing over and recombination during meiosis for an autosomal chromosome. After duplication the chromatids cross over, which generates gametes with alternating segments of the grand-paternal and grand-maternal chromosomes. Tightly-linked loci (markers 1 and 2) are often segregated together whereas weakly linked loci (markers 2 and 3) are segregated nearly independently of each other.

Two loci located near each other on the same chromosome will tend to be passed together from parent to child, and the closer the loci are to each other the less likely a recombination event will separate them. The simultaneous segregation of loci is further complicated by the fact that the chiasmata do not occur entirely at random along the chromatids. Crossovers become more frequent away from the centromere and two chiasmata seldom occur close to each other. However, when two chromosomal regions are very far apart on the same chromosome, the average number of crossover events between them equals the number of non-crossover events. Thus the gametes produced end up having all combinations of the two regions in equal proportions. Furthermore, the average number of crossovers in female meioses is greater than in male meioses, giving different recombination patterns from maternally and paternally derived chromosomes, and adding to the complexity of creating genetic linkage maps.

Since crossovers only occur between non-sister chromatids, the number of chiasmata seen for a pair of sister chromatids equals the total number of chiasmata in all four strands (e.g. a total of two crossovers occur on Figure 1. These two crossovers are seen as two crossovers for the two light chromosomes and as two crossovers for the dark chromosomes). If the total number of chiasmata is odd, one strand
of a pair of sister chromatids must contain an odd number of crossovers, while the other strand must contain an even number. When the total number of chiasmata is even, both strands of the pair of sister chromatids contain either an even or an odd number of crossovers. Hence, a single randomly selected homologous strand is equally likely to be part of an even or an odd number of crossovers as long as at least one crossover occurs for the four strands (assuming the chromatids participating in the crossover are selected uniformly and independently from chiasma to chiasma).

**Linkage**

The **recombination fraction** between two loci \( A \) and \( B \) will be denoted \( \theta_{AB} \) and is defined as the probability of having a recombination (an odd number of crossovers) between the two loci (i.e., the two alleles at the two marker loci have different ancestral origin).

Continuing the arguments from the previous section, the probability of a recombination between two marker loci is \( \frac{1}{2} \) as long as at least one crossover occurs between them. This restricts the recombination fraction between two marker loci on the same chromosome to the interval \([0, \frac{1}{2}]\), with a recombination fraction of \( \frac{1}{2} \) meaning that the marker information at one locus is independent of the marker information at the other locus. The recombination fraction between two marker loci on different chromosomes is defined as \( \frac{1}{2} \) as chromosomes are segregated independently. In other words, the marker information from two loci on different chromosomes is independent.

Mather’s formula relates the recombination fraction to the total number of chiasmata. Let \( \theta_{AB} \) be the recombination fraction between locus \( A \) and locus \( B \) and let \( N_{[A,B]} \) be the total number of chiasmata between the two loci. The formula can be derived as follows:

\[
\theta_{AB} = P(\text{recombination between locus } A \text{ and } B) \\
= P(\text{recomb. between } A \text{ and } B | N_{[A,B]} = 0)P(N_{[A,B]} = 0) \\
+ P(\text{recomb. between } A \text{ and } B | N_{[A,B]} > 0)P(N_{[A,B]} > 0) \\
= \frac{1}{2}P(N_{[A,B]} > 0) = \frac{1}{2}(1 - P(N_{[A,B]} = 0)) \tag{1}
\]

The last line is a direct result of the arguments presented above.

Two loci on a chromosome are said to be **linked** (segregating together) if the recombination fraction between them is less than \( \frac{1}{2} \). They are completely linked when the recombination fraction is 0 and they are **unlinked** if the recombination fraction is \( \frac{1}{2} \).

**Linkage analysis**

In linkage analysis we seek to determine if two loci (or a locus and a trait) are segregating together. If a locus and a trait are segregating together, we have identified chromosomal position that — even if the locus itself does not influence the trait — is in the vicinity of gene(s) that influence the trait. If two loci we found to be linked we know that they must be physically close on a chromosome (loci on different chromosomes are assumed to segregate independently).

Thus, we are interested in the one-sided hypothesis of no linkage:

\[
H_0 : \theta_{AB} = \frac{1}{2} \text{ vs. } H_A : \theta_{AB} < \frac{1}{2}. \tag{2}
\]

If we reject the hypothesis we say that \( A \) and \( B \) are linked.

1. **Back cross design.** In a back cross design we know the original two different inbred lines found in the parents. When we cross the \( F_1 \) individual with one of the parents we only observe recombinations on one gamete (the one from the \( F_1 \) individual) and since the two original lines were inbred we can see if an allele if from one of the parental lines or from the lines of other parent.

That means that in a back cross design we can easily determine if a recombination has taken place and we can simply count the number of individuals (back crosses) that have recombinations.

The number of recombinations can be modeled as a binomial distribution

\[
Y \sim \text{binom}(n, \theta_{AB})
\]
and we have that \( \hat{\theta}_{AB} = y/n \wedge 1/2 \) and that the likelihood ratio test for the no linkage hypothesis (2) is

\[
\left( \frac{y}{n} \right)^y \left( \frac{1 - y}{n} \right)^{n-y} \frac{1}{2}
\]

if \( y/n \leq 1/2 \). The log likelihood ratio test statistic has an asymptotic distribution that is a 50:50 mixture of a \( \chi^2_1 \) distribution and an atom in 0 because we are testing a parameter on the boundary of the parameter space.

For example, assume we have data on two loci for 100 back crosses. We also know that the original inbred lines are \( AB \) (i.e., allele \( A \) at the first locus and allele \( B \) at the other locus) and \( ab \). We want to examine if the two loci segregate together. The observed data are

<table>
<thead>
<tr>
<th>First locus</th>
<th>Second locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>33</td>
</tr>
<tr>
<td>( a )</td>
<td>21</td>
</tr>
</tbody>
</table>

We can see, that there is 33 + 32 = 65 non-recombinations and 21 + 14 = 35 recombinations. Thus, \( \hat{\theta} = 35/100 = 0.35 \) and twice the likelihood ratio test statistics is 9.14. We clearly reject the null hypothesis and conclude that the two loci are linked.

2. **Nuclear families.** The computations for a nuclear family (a family with father, mother and \( k \) offspring) are nowhere as simple as for the back crosses. The problem is, that in order to determine if a recombination has occurred we need to know the phase of all genotype, i.e., we need to know exactly which alleles are inherited from the father and which alleles are inherited from the mother and we need to know the haplotypes of both parents. In the back cross case we did not have this problem because we only had one gamete where recombination could occur and because the original inbred lines were different.

In some situations with nuclear pedigrees it is possible to determine which alleles were inherited from each parent, but in general that is not possible. In these situations we need to use combinatorics and go through all possible combinations of ordered genotypes (phase-known genotypes) that are consistent with the observed data and — for each combination — determine in which offspring we see recombinations and in which there are no recombinations.

Let us illustrate this with an example. In the pedigree from figure 2 we see that mother is non-informative for linkage as she is doubly homozygous. That also makes it easy to determine the haplotypes for the 5 offspring as they all must have inherited a 1A haplotype.

We can only determine if the 5 offspring have had a recombination or not if we know the two haplotypes (phases) of the father. If the father is 1A|2a then two of the offspring are recombinant and if the father is 1a|2A then three of the offspring are recombinant. A priori we have no knowledge about the parental haplotypes and if the marker loci are in linkage equilibrium the two haplotype combinations are equally likely.

When we integrate out the paternal phase distribution the likelihood ratio test statistic of no linkage for the pedigree becomes

\[
\frac{1}{2} \left[ \theta^2 (1 - \theta^3) + \theta^3 (1 - \theta)^2 \right]
\]
Figure 3: There are four possible recombinant classes, I–IV, for three marker loci A, B and C with known order. “No recombination” and “recombination” correspond to an even or odd number of crossovers between two markers. \( \gamma_{00}, \ldots, \gamma_{11} \) are the probabilities of the different recombinant classes and \( \theta_{AB} \) and \( \theta_{BC} \) are the recombination fractions.

where we have used that conditional on the parental haplotype configuration we can determine the recombinations in the offspring, and know that the number of recombinations follow a binomial distribution.

Basically, we need to go through all possible ordered combinations for all loci for all individuals. Many of the combinations can be ruled out because they are inconsistent with Mendels’ laws and receive weight 0, and the remaining combinations are all equally likely if we have no missing data. If parental genotypes are missing we will have to fill in genotypes according to the population genotype frequencies and weigh each genotype configuration with the probability of the parental haplotype configuration.

Three point linkage

In this section we consider the case where we have information on three marker loci. This enables us to estimate any interaction in recombination frequencies between neighboring intervals and to determine the order in which the markers are located (see the section below on genetic maps).

For an ordered set of markers \( A — B — C \), a recombination between \( A \) and \( C \) occurs only if there is an odd number of crossovers in exactly one of the two intervals (see also Figure 3 for illustration). By definition

\[
\theta_{AC} = \theta_{AB} - \theta_{ABC} - \theta_{AB} + \theta_{BC} - 2\theta_{ABC}
\]

with \( \theta_{ABC} \) being the probability of observing an odd number of crossovers in both the interval between \( A \) and \( B \) as well as in the interval between \( B \) and \( C \) (a double recombination).

In the three-locus example, the recombination fractions \( \theta_{AB}, \theta_{BC} \) and \( \theta_{AC} \) can also be parameterized by the frequency of the four possible recombinant classes as shown in Figure 3. Each meiotic product will belong to exactly one of the four recombinant classes as there will be either an even or an odd number of crossovers between \( A \) and \( B \), and between \( B \) and \( C \) with frequencies \( \gamma_{00}, \ldots, \gamma_{11} \). Figure 3 also shows the case of double crossover (class IV) where there is a recombination between marker \( A \) and marker \( B \) and recombination between marker \( B \) and marker \( C \), but no observable recombination between marker \( A \) and marker \( C \). In many situations, this parameterization is easier to work with and it simplifies calculations for multiple loci.

The relationship between the frequency of the recombinant classes and the recombination fractions can be written as

\[
\begin{align*}
\theta_{AB} &= \gamma_{11} + \gamma_{10} \\
\theta_{BC} &= \gamma_{11} + \gamma_{01} \\
\theta_{AC} &= \gamma_{10} + \gamma_{01}
\end{align*}
\]

with the restrictions that \( \gamma_{00} + \gamma_{10} + \gamma_{01} + \gamma_{11} = 1 \) and that \( \gamma_{10} \leq \frac{1}{2}, \gamma_{01} \leq \frac{1}{2} \) and \( \gamma_{11} \leq \frac{1}{2} \) since the \( \gamma \)'s are bounded upwards by \( \frac{1}{2} \).

Notice that the recombinant classes correspond to the haplotypes. That means that if we want to estimate the three recombination frequencies, \( \theta_{AB}, \theta_{BC} \) and \( \theta_{AC} \), we could just estimate frequency of each recombinant class. Since the recombinant classes correspond to haplotypes we can estimate the
haplotype frequencies in exactly the same way as we did for the recombination frequency in the case with two marker loci. The only difference is that instead of the binomial distribution we use a multinomial distribution because there are 4 recombinant classes. The method, however, still requires us to go through all possible phase combinations and weight the observed phase combination according to the likelihood of the pedigree.

**General likelihood of pedigree**

Here we will formulate the likelihood of a pedigree. The likelihood will provide the weights needed to estimate the frequencies of the recombinant classes (haplotypes) that are needed to determine two-, three- and multipoint linkage.

Let \( X_i \) be the observed genotype (denoted the phenotype in the following since it is the appearance of the true genotype) of individual \( i \) at a marker locus and let \( G_i \) be the true, ordered genotype at that locus. \( P(X_i|G_i) \) denotes the probability of observing phenotype \( X_i \) when the true genotype is \( G_i \) and \( P(G_i) \) denotes the probability of a random individual from the population having the true genotype \( G_i \). Pedigree founders are assumed to be independent and since no-founder genotypes only depend on the parental genotypes, the likelihood of a single pedigree with \( n \) individuals at a single marker locus can be written as

\[
L(X) = \sum_{G_1 \in \mathcal{G}} \cdots \sum_{G_n \in \mathcal{G}} \prod_{i=1}^{n} P(X_i|G_i) \prod_{j \text{ founder}} P(G_j) \prod_{o \text{ offspring}} P(G_o|G_{o_1}, G_{o_m})
\]

where \( P(G_o|G_{o_1}, G_{o_m}) \) is the probability that a genotype \( G_o \) is transmitted to a non-founder \( o \) having a father with genotype \( G_{o_1} \) and a mother with genotype \( G_{o_m} \). The \( \mathcal{G}_1, \ldots, \mathcal{G}_n \) are the sets of possible ordered marker genotypes for the \( n \) individuals. In essence, the likelihood examines all possible combinations of genotypes and only includes the ones that are compatible with the observed data.

For a marker with \( A \) alleles each set \( \mathcal{G} \) contains \( A^n \) elements making the direct evaluation of (5) computationally infeasible. Several improvements exist to reduce the dimensionality of the problem and to make likelihood evaluations feasible. If the founder population at the marker is assumed to be in Hardy-Weinberg equilibrium the \( P(G_i) \) probability is just the product of the allele frequencies for the alleles determined by \( G_i \). This reduces the number of parameters used for specifying the distribution of the genotypes in the founder population from \( 2A - 1 \) to \( A - 1 \) (implicitly assuming that genotype frequencies do not depend on the order of the alleles).

Since we are only considering ordered genotypes, the transmission probabilities \( P(G_o|G_{o_1}, G_{o_m}) \) partitions into the product \( P(G_o|G_{o_1})P(G_{o_m}|G_{o_1}) \), as parents segregate alleles independently of each other. \( P(G_o|G_{o_1}) \) is the probability of an offspring receiving allele \( G_o \) from the father given the fathers genotype, and \( P(G_{o_m}|G_{o_1}) \) is defined similarly. The transmission probabilities are zero if the allele could not have come from the parent. If the allele of the offspring is present in the parent, the transmission probability is \( \frac{1}{2} \) when the parent is heterozygous and 1 if the parent is homozygous.

The observed phenotype always equals the true genotype when the markers are codominant and are observed without error. This can be written as

\[
P(X_i|G_i) = \begin{cases} 
1 & \text{if } X_i = G_i \\
0 & \text{otherwise} 
\end{cases}
\]

and means we can remove all the genotypes from the sets of possible ordered genotypes \( \mathcal{G}_1, \ldots, \mathcal{G}_n \) that are not consistent with the observed marker information for the different individuals. This simple step makes a cardinal reduction in the computational demands of the likelihood. The size of each set \( \mathcal{G} \) is reduced to one or two elements if an individual is genotyped (one element if the genotyped individual is homozygous at the locus and two elements if the individual is heterozygous). The genotype-phenotype penetrance probability (6) is easily modified to handle non-codominant markers. There are 9 true, ordered genotypes for the ABO blood group marker (00, A0, 0A, B0, 0B, AA, AB, BA, and BB) but only 4 observable phenotypes (0, A, B, and AB). This can be handled by defining \( P(X_i|G_i) \) accordingly and will still result in making the computation of the likelihood more tractable. Untyped individuals are handled by not imposing restrictions on the possible ordered genotypes for that individual.

The likelihood in (5) only assumes a single marker locus, but with minor modifications it can handle multiple loci. This is done by considering observed and true haplotypes instead of genotypes. Let \( X_i \) be the observed pairs of haplotypes for individual \( i \) in a pedigree and let \( G_i \) be the pair of true haplotypes.
The multilocus generalization of the likelihood (5) for a single pedigree is

\[ L(X) = \sum_{G_1 \in G_1} \cdots \sum_{G_n \in G_n} \prod_{i=1}^{n} P(X_i | G_i) \prod_{j \text{ founder}} P(G_j) \prod_{o \text{ non-founder}} P(G_o | G_{of}, G_{om}) \]  

(7)

where \( P(G_j) \) is the frequency of the haplotype and where \( P(X_i | G_i) \) is the multivariate penetrance function. Both the haplotype frequencies and the penetrance probabilities simplify to the product of the marker-specific allele frequencies or penetrance functions, respectively, if the marker loci are in linkage disequilibrium. The largest difference from (5) to (7) is in the transmission probabilities, \( P(G_o | G_{of}, G_{om}) \), as these not only depend on parental haplotypes, but also on the frequency of the recombinant classes that determine any crossovers between loci. In theory, the multivariate likelihood can handle any number of marker loci simply by specifying the haplotype and recombinant class frequencies. For any real dataset, however, the number of possible combinations required to evaluate the likelihood increases exponentially with the number of loci, and it may be necessary to make several assumptions to simplify the likelihood calculation.

One can calculate the exact likelihood and estimate the frequencies of the recombinant classes simply by counting haplotypes when it is possible to determine recombinations and non-recombinations without errors. This is only possible when the genetic data are phase known, all individuals are fully informative, and there are no missing data; these are rare qualities of non-inbred, random mating populations. The phase of the founders is hardly ever unknown, the individuals are not fully informative and missing data occur frequently. If data are not fully informative or there is missing data, it is necessary to integrate over all possible ordered genotypes for the non-informative individuals as in (5).

Using linkage to construct genetic maps

Genetic linkage maps play an important role in statistical genetics as well as in genetic epidemiology; they are essential tools for locating genes with traditional linkage analysis or QTL mapping. For some organisms, such as the human and the mouse, the number of available genetic markers has increased so dramatically in recent years that it is possible to create very dense genetic maps for fine-scale mapping of disease genes.

Genetic maps show the ordering of loci along chromosomes and the relative distances between them, providing “landmarks” for positioning putative trait-influencing genes. Besides non-coding genetic markers, these landmarks can include known genes and without the landmarks, it is virtually impossible to determine the location of a trait-influencing.

Creation of multilocus linkage maps began with a paper by Sturtevant (1913), showing the (correct) order and distance between six sex-linked genes in Drosophila. The construction of genetic linkage maps involves the following three parts:

1. Classifying genetic markers in linkage groups (chromosomes).
2. Determining the order of loci (markers) along each chromosome.
3. Estimating the relative distance between loci in each linkage group.

The three steps outlined above need not be done separately, often steps 2 and 3 are combined, and the order and the distances are estimated simultaneously.

The problems of determining linkage groups and ordering markers along chromosomes for humans have more or less disappeared after the Human Genome Project (HUGO) published the first draft of the sequence of the human genome (Lander et al., 2001; Venter et al., 2001). Linkage groups and marker orders can be determined simply by comparing the marker sequences to the human sequence, which reveals the correct chromosome and position of each genetic marker. When the human genome has been fully sequenced these two problems will disappear altogether but it will still be necessary to estimate the recombination fractions between the markers. Also, the methods are still relevant for other species besides humans.

While the Human Genome Project and other genome projects can provide information on the number of base pairs between the markers, this information cannot be used to determine the recombination frequencies. Since crossovers occur with varying frequencies along the chromosomes, the relationship between recombination fraction and number of base pairs is highly dependent on the part of the chromosome being analyzed. Furthermore, the physical distance between two markers can vary greatly with for example ethnicity.
Determining linkage groups

Traditionally, linkage groups are identified by comparing the recombination fractions of all possible two-point linkage analyses. Two markers are said to be in the same linkage group if they are linked.

For the purpose of determining linkage groups, linkage between markers is considered transitive. That is, if marker $A$ is linked to marker $B$, and if $B$ is linked to $C$, then $A$, $B$, and $C$ will be included in the same linkage group, although markers $A$ and $C$ are not necessarily linked.

By testing for linkage between $m$ markers, a total of $\frac{m(m-1)}{2}$ hypotheses are being tested. The significance level of these linkage tests is usually set very low to minimize the risk of collapsing linkage groups (i.e., making a statistical type I error) due to multiple comparisons. For humans, a $p$-value less than 0.0001 is commonly used to indicate linkage, but the threshold for linkage obviously depends on density of the genetic markers used as well as the species examined.

Lincoln et al. (1993) suggest adding another criterion before including a marker in a linkage group. Besides having a LOD-score greater than three, the estimated distance between the two markers should be no greater than a certain threshold to further reduce the risk of including a marker in a linkage group due to pure chance. This corresponds to replacing the hypothesis of no linkage (2) with the hypothesis $H_0 : \theta_{AB} \geq \tilde{c}$ vs. $H_A : \theta_{AB} < \tilde{c}$, where $\tilde{c}$ is a constant depending on the maximum allowed distance through the mapping function used (e.g., $\tilde{c} = .40$).

The threshold used for finding linkage groups should be based on the total length of the genetic map of the species examined, the number of markers as well as the density of the markers, if known. A sparse set of genetic markers can give too many linkage groups because markers from the same group may be so far apart that they show no linkage between them. Using a dense set of markers, however, can result in too few linkage groups, because the large number of pairwise linkage tests gives increased risk of seeing two markers from different linkage groups showing significant linkage by chance.

Ordering markers

Once the genetic markers have been assigned to chromosomes, the next step is to determine the order of the markers along each chromosome. In theory, the best order of the markers is the one having the largest likelihood, but calculating the likelihood for all the possible marker orders is not always possible in practice. For $m$ markers on a chromosome, there are $\frac{m!}{2}$ possible orders and even for a moderate number of markers, the computational burden of going through all possible orders becomes too large. Moreover, even when a specific order is given, calculating the likelihood of the data can be a non-trivial computational task as mentioned in section.

The problem of finding the most likely order of $m$ markers is analogous to the classical “traveling salesman problem” and the solutions to that problem can be applied directly to the problem of ordering markers. Unfortunately, there is no simple analytical solution to this problem, and it is necessary to either reduce the problem and/or use numerical search methods to find a reasonable solution (Lawler et al., 1985; Groetschel and Holland, 1991).

Three different approaches are commonly used for determining the order of the $m$ marker loci.

- The first approach is to reduce the size of the problem by only considering a small subset of the $\frac{m!}{2}$ possible orders, and selecting the order from the subset having the largest likelihood. Obviously, this approach replaces the original problem with the problem of selecting a small subset that has a high probability of containing the order with the overall highest likelihood of all possible orders. One way to select a smaller subset to analyze is to exclude highly improbable marker orders (based on a likelihood-ratio test).

A set of probable marker orders can be generated by starting with a map containing only two marker loci and then adding new marker loci one at a time. The first added marker can be placed in three positions (to the left of the two initial markers, between them and to the right of them), the next in four positions etc. giving a $\sum_{k=3}^{m} k = \frac{(m-2)(m+3)}{2}$ evaluations of marker order if the order in which to add markers is fixed. That is not generally the case so markers are added according to their increase in likelihood (i.e., the two initial markers are the pair of markers having the highest likelihood. Each of the remaining markers are examined at the three possible positions and the marker/position combination having highest likelihood is used for selecting and placing marker three, etc.) The total number of likelihood evaluations (which includes different numbers of markers) is $\frac{1}{2} \sum_{k=1}^{m} k(m-k+1)$.

When there is no unique best marker/position combination (i.e., if no new marker placement is significantly better than all other possible marker placements using a standard likelihood ratio
test), all of the marker/position combinations with a high likelihood are examined when placing the next marker.

A variation of this approach is used in the computer program CRI-MAP (Green et al., 1990). However, instead of estimating pairwise recombination fractions from all individuals, CRI-MAP disregards all non-informative genotypes and uses only information from the heterozygous genotypes. Simulation studies have shown that the estimated recombination fractions from CRI-MAP are almost identical to the estimates from a full likelihood program (Goldgar et al., 1989). The advantage of this approach is that allele frequencies are not needed to evaluate potential genotype frequencies because ambiguous genotypes are discarded and likelihood evaluation is simple since only informative meioses are left.

- A second approach is to estimate all \( \frac{m(m-1)}{2} \) possible pairwise recombination fractions, and then derive the order of the marker loci from these estimates (Buetow and Chakravarti, 1987). This approach, known as seriation, generates \( m \) possible different marker orders, and these orders should be reduced to a single order using ranks or a goodness-of-fit test (Gelfand, 1971). The \( m \) possible orders are made by starting with a single marker, and then adding the remaining markers in the order of their closeness (measured in recombination fraction) to the starting marker. The markers are added to the end of the existing chain, where the recombination fraction between the new marker and the marker at the endpoint is smallest. Ties are resolved by examining more markers as in the approach discussed above.

Several variations on the seriation approach exist. Doerge (1993) suggests a method (called Rapid Chain Delineation) that removes the need to average the \( m \) possible markers by generating a single marker order. Rapid Chain Delineation uses the two most closely linked markers for initializing the seriation algorithm. Before adding a new marker to the current marker order it is determined if the likelihood of the current marker order can be improved upon by permuting any adjacent marker loci. Intuitively this is reasonable as the most likely error is to get the order of two closely linked markers wrong.

- The final approach is the use of simulated annealing. Simulated annealing is a Monte Carlo approach for minimizing functions, that is motivated from thermodynamics, where a liquid reaches a minimum energy state when cooled slowly.

Let \( E_i \) denote the total energy of the system when in state \( i \). Since the energy will be minimized in simulated annealing, one of the following relationships can be used to maximize the likelihood

\[
E_i = \begin{cases} 
- \log L(i) & \\
1 / \log L(i) 
\end{cases}
\]

(8)

where \( i \) is a marker order. The exact choice of relationship between the energy function and the likelihood in the above formula is not important. However, the choice influences the critical temperature levels and the rate at which the temperature should be lowered. The major drawback with simulated annealing compared to the two other approaches is that it requires many evaluations of the likelihood of all \( m \) markers.

To apply simulated annealing to marker ordering, the system is initialized with a particular locus order, \( i \). A new order, \( j \), is proposed by a random displacement, and the new order is accepted if the energy of the new order is lower than the energy of the current order, i.e., if \( E_j \leq E_i \). If \( E_j > E_i \), a Metropolis step is applied and the new state, \( j \) is accepted with probability \( \exp(-\frac{E_j - E_i}{T}) \), where \( T > 0 \) represents temperature. This procedure allows the system to move consistently toward lower energy states, yet still jump out of local minima due to the probabilistic acceptance of some upward moves. The annealing process should stop in a state with minimum or nearly minimum energy if the temperature is decreased gradually toward zero.

There are \( \frac{m!}{2} \) possible locus orders for \( m \) markers. The displacement consists of block inversions of markers from the current order. For example, if the current state is 1–2–3–4–5–6–7, a random block, such as 4–5–6, is selected and inverted to yield the new order 1–2–3–6–5–4–7. The block to invert is chosen uniformly from the \( \frac{m(m-1)}{2} \) possible transitions.

Despite the fact that the convergence rate of simulated annealing can be slow, the method has proven very effective in finding solutions to the traveling salesman problem and has also been used for ordering markers (Weeks and Lange, 1987).
All three approaches estimate an approximate order, but none of them are guaranteed to give the best order in a realistic amount of time. The first two approaches are reasonably fast, even for a large number of marker loci, whereas simulated annealing can be slow due to many time-consuming likelihood evaluations. A reasonable strategy for ordering marker loci would therefore be to start with one of the faster algorithms to get a good starting guess of the orders and then try to improve the resulting order by using simulated annealing. When the best order has been approximated, it is a good idea to permute the order of neighboring marker loci to see whether the likelihood of the permuted order increases.

Other methods for solving the “traveling salesman problem” can also be applied to finding the marker order. One quite effective method involves an artificial ant colony (Dorigo and Gambardella, 1997). It combines some of the ideas of simulated annealing with the ideas of seriation.
References


